SAR 2020/2021 Laboratorium 2

13-14.10.2020

2.1 (model regresji jednokrotnej, rozkład QR)

Wygenerować chmurę punktów w następujący sposób. Niech x_i będą punktami z zakresu $[0, 10]$ odległymi o 0.1, natomiast $y_i = x_i + \epsilon_i$, gdzie ϵ_i to zmienne losowe z rozkładu normalnego o średniej 0 i odchyleniu standardowym $\sigma = 3$. Wykonaj wykres rozproszenia (x_i, y_i).

a) Oblicz współczynnik korelacji próbkowej korzystając z definicji oraz funkcji `cor()`.

b) Oblicz współczynniki prostej MNK korzystając z funkcji `lm()`. Następnie policz współczynniki na cztery inne sposoby:
- jako minimum funkcji kryterialnej $Q((\beta_0, \beta_1)) = \sum_{i=1}^{n} ||y_i - \beta_0 - \beta_1 x_i||^2$ (funkcja `optim()`),
- z definicji wyznaczając macierz eksperymentu X,
- przekształcając równanie normalne (czyli $X'X\beta = X'Y$) korzystając z rozkładu QR (funkcja `qr()`) dla macierzy eksperymentu $X = QR$, a potem rozwiązując układ równań liniowych (na laboratoriach można użyć funkcji `solve()`, w domu można napisać „od zera"), - optymalizując funkcję log-wiarogodności.

c) Nanieść otrzymaną prostą MNK na wykres rozproszenia.

d) Powtórzyć procedurę z punktów (a)-(c) dla $\sigma = 0.5$, $\sigma = 5$. Co można powiedzieć o wartości współczynnika korelacji próbkowej oraz współczynnikach prostej MNK?

2.2 (model regresji jednokrotnej, współczynnik determinacji)

W bibliotece `MASS` znajduje się zbiór danych *hills* dotyczących biegów przełajowych, które odbyły się w Szkocji w 1984 roku. Zawiera on trzy zmienne:

time - rekordowy czas pokonania trasy (w minutach),

dist - długość trasy w milach (na mapie),

climb - całkowita różnica poziomów do pokonania na trasie (w stopach).

a) Porównać (wyświetlając w jednym oknie) wykresy rozproszenia *time* od *dist* i *time* od *climb*. Obliczyć współczynniki korelacji *time* i *dist* oraz *time* i *climb*.

b) Dopasować proste MNK opisujące zależność *time* od *dist* oraz *time* od *climb*. Nanieść dopasowane proste na wykresy rozproszenia. Obliczyć R^2 dla otrzymanych modeli posługując się trzema metodami:
- z definicji, wyznaczając wartości SST, SSR i SSE,
- korzystając z funkcji `cor()` (R^2 to kwadrat korelacji pomiędzy zmienną objaśnianą i objaśniającą),
- korzystając z funkcji `summary()`.

c) Jaki rekordowy czas pokonania trasy o długości 15 mil przewidzimy posługując się prostą MNK? Policzyć to na dwa sposoby:
- korzystając z funkcji `predict()` i odpowiedniego modelu,
- korzystając z wyestymowanych współczynników.

`Adjusted R-squared` jest estymatorem ρ^2, gdzie ρ jest korelacją wielokrotną. Zależność pomiędzy R^2 i R^2_{adj} to $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$.

2.3 (anscombe quartet)

Wczytać zbiór *anscombe_quartet.txt*.

a) Do każdej z czterech par zmiennych dopasować prostą MNK.

b) Porównać otrzymane współczynniki dopasowanych prostych MNK, współczynniki R^2 i współczynniki korelacji.

c) W jednym oknie sporządzić 4 wykresy rozrzutu Y_i od X_i, $i = 1, 2, 3, 4$. W którym przypadku można mówić o przybliżonej zależności liniowej y od x?

2.4 (model regresji wielokrotnej)

Plik *realest.txt* zawiera następujące dane na temat domów na przedmieściach Chicago: cena domu (*Price*), liczba sypialni (*Bedroom*), powierzchnia w stopach kwadratowych (*Space*), liczba pokoi (*Room*), szerokość frontu działki w stopach (*Lot*), roczny podatek od nieruchomości (*Tax*), liczba łazienek (*Bathroom*), liczba miejsc parkingowych w garażu (*Garage*) i stan domu (*Condition*, 0 - dobry, 1 - wymaga remontu). Dopasować liniowy model regresji opisujący zależność ceny domu od pozostałych zmiennych w zbiorze.

a) Wyznacz macierz eksperymentu.

b) Oblicz estymatory parametrów z definicji i porównaj z wartościami obliczonymi przy użyciu funkcji `coef()`. Oblicz SST, SSR, SSE i współczynnik determinacji z definicji.

c) Jaki wpływ na cenę ma zwiększenie liczby sypialni o 1, kiedy wartości wszystkich pozostałych zmiennych objaśniających są ustalone? Znaleźć uzasadnienie tego pozornie błędnego wyniku. Porównać ten wynik z wynikiem otrzymanym dla modelu linowego opisującego zależność ceny domu jedynie od liczby sypialni.

d) Masz dom w tej okolicy, w dobrym stanie, z 3 sypialniami, o powierzchni 1500 stóp kwadratowych, z 8 pokojami, 40 stopami szerokości działki, 2 łazienkami, 1 miejscem w garażu i podatkiem w wysokości 1000 dolarów. Za ile spodziewasz się go sprzedać? Wykonaj predykcje korzystając z definicji oraz funkcji `predict()`.

e) Oblicz estymator wariancji błędów korzystając z definicji oraz funkcji `summary()`.

f) Oblicz estymatory błędów standardowych współczynników korzystając z definicji oraz funkcji `summary()`.